

# Strong consistency of the maximum likelihood estimator for finite mixtures of location-scale distributions when penalty is imposed on the ratios of the scale parameters

Kentaro Tanaka\*

February 8, 2008

## Abstract

In finite mixtures of location-scale distributions, if there is no constraint or penalty on the parameters, then the maximum likelihood estimator does not exist because the likelihood is unbounded. To avoid this problem, we consider a penalized likelihood, where the penalty is a function of the minimum of the ratios of the scale parameters and the sample size. It is shown that the penalized maximum likelihood estimator is strongly consistent. We also analyze the consistency of a penalized maximum likelihood estimator where the penalty is imposed on the scale parameters themselves.

*Key words and phrases:* penalized likelihood; unboundedness of likelihood

## 1 Introduction

In this paper, we prove the strong consistency of a penalized maximum likelihood estimate for finite mixtures of univariate location-scale distributions generalizing the results in Ciuperca, Ridolfi, and Idier (2003). As a special case of this result, we solve an open problem posed by Hathaway (1985).

As stated in Day (1969), because the likelihood function for finite mixtures of location-scale distributions is unbounded, the maximum likelihood estimator does not exist. To see that, we consider a simple case that the model consists of mixtures of two normal distributions  $\alpha_1\phi(x; \mu_1, \sigma_1) + \alpha_2\phi(x; \mu_2, \sigma_2)$  and assume that we obtain an i.i.d. sample  $X_1, X_2, \dots, X_n$  from the true distribution. If we set  $\mu_1 = X_1$  and  $\sigma_1 \rightarrow 0$ , then the likelihood tends to infinity as  $\sigma_1$  goes to zero. Hence the likelihood function is unbounded.

---

\*Department of Industrial Engineering and Management, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552 JAPAN, E-mail: tanaka.k.al@m.titech.ac.jp

A straightforward approach to this problem is to bound the minimum of the variances of the components from below by a positive constant. By using theorem 6 in Redner (1981), we can show that the maximum likelihood estimator restricted to a compact subset of the parameter space is strongly consistent if the subset contains the true parameter.

Another approach is penalized maximum likelihood estimation. However, if the penalty is not appropriate, then the likelihood function is unbounded. Ciuperca, Ridolfi, and Idier (2003) considered the case that the penalties are imposed on the variances themselves and proved the consistency of the penalized maximum likelihood estimator. The results given in Ciuperca, Ridolfi, and Idier (2003) are very useful for estimating the parameters of mixture of normal densities because the assumptions for the penalty are easy to check and the implementation of their method is also easy. In this paper, we extend their consistency result to the case that the components of mixtures are not normal densities and the penalty depends on the sample size  $n$ .

In normal mixture distributions, Hathaway (1985) considered the following constraints to avoid the divergence of the likelihood.

$$\min_{m,m'} \frac{\sigma_m}{\sigma_{m'}} \geq b \quad (1.0.1)$$

This bounds the minimum of the ratios of the variances of the components by a constant. He showed that the strong consistency of the maximum likelihood estimator holds if the true distribution satisfies the constraint represented by equation (1.0.1). Intuitively, a stronger constraint must be enforced for a smaller sample size to avoid the divergence of the likelihood, because a component with a very small variance can only have a large contribution to a single observation at most. Therefore, it seems that the constraint under which the consistency holds can be weakened as the sample size increases. This intuition leads to the following two questions:

- Is it possible to let  $b$  decrease to zero as the sample size  $n$  increases to infinity while maintaining consistency?
- If it is possible, then at what rate can  $b$  be decreased to zero?

These questions are mentioned in Hathaway (1985), McLachlan and Peel (2000), and treated as unsolved problems.

This topic is closely related to a sieve method. (See Grenander (1981) and Geman and Hwang (1982). ) For normal mixture distributions, the convergence rate of the maximum likelihood estimator based on sieve method is studied in Genovese and Wasserman (2000) and Ghosal and van der Vaart (2001). In Tanaka and Takemura (2006), for mixtures of location-scale distributions, we showed the strong consistency of the maximum likelihood estimator in the case that the scale parameters themselves are bounded from below by  $c_n = e^{-n^d}$ , ( $0 < d < 1$ ). However, we could not solve the original questions when constraints are imposed on the minimum of the ratios of the variances of the components.

In this paper, we solve the questions treated above in a more general and unified framework. For mixtures of location-scale distributions, we consider a penalized likelihood, where the penalty is a function of the minimum of the ratios of the scale parameters

and the sample size  $n$ . The effect of the penalty becomes stronger as the minimum of the ratios of the scale parameters decreases to zero. Note that the penalty can depend on the sample size  $n$ . We can weaken the effect of the penalty as the sample size  $n$  increases to infinity. In Theorem 1, we show that the consistency holds for the penalized maximum likelihood estimator. In Corollary 1, the solutions to the questions mentioned in Hathaway (1985), McLachlan and Peel (2000) are obtained as special cases of Theorem 1. We also analyze the consistency of a penalized maximum likelihood estimator in which the penalties are imposed on the scale parameters themselves. The result obtained in Theorem 2 is a generalization of Corollary 1 of Ciuperca, Ridolfi, and Idier (2003).

Throughout this paper, we assume that the true distribution is a mixture of location-scale distributions and the number of components of the true distribution is known.

The organization of this paper is as follows. Section 2 describes notation and regularity conditions. The main results are stated in section 3. Section 4 is devoted to the proofs. We end this paper by concluding remarks in section 5.

## 2 Preliminaries

### 2.1 Notation

Mixture of  $M$  location-scale densities are written in the form

$$f(x; \theta) \equiv \sum_{m=1}^M \alpha_m f_m(x; \mu_m, \sigma_m).$$

The mixing weights  $\alpha_1, \dots, \alpha_M$  have to satisfy  $\alpha_m \geq 0$ ,  $\sum_{m=1}^M \alpha_m = 1$ . We assume that the components  $f_1(x; \mu_1, \sigma_1), \dots, f_M(x; \mu_M, \sigma_M)$  are location-scale densities i.e. they satisfy

$$f_m(x; \mu_m, \sigma_m) = \frac{1}{\sigma_m} f_m\left(\frac{x - \mu_m}{\sigma_m}; 0, 1\right) \quad , \quad 1 \leq m \leq M,$$

where  $\mu_m$  and  $\sigma_m$  are location parameters and scale parameters respectively. We abbreviate  $(\alpha_1, \mu_1, \sigma_1, \dots, \alpha_M, \mu_M, \sigma_M)$  as  $\theta$ , and  $(\mu_m, \sigma_m)$  as  $\theta_m$ . We denote the true parameter by  $\theta_0$ .

Let  $\Omega_m \equiv \{(\mu_m, \sigma_m) \mid \mu_m \in \mathbb{R}, \sigma_m \in (0, \infty)\}$  denote the parameter space of the  $m$ -th component. Then the entire parameter space  $\Theta$  can be represented as

$$\Theta \equiv \{(\alpha_1, \dots, \alpha_M) \mid \sum_{m=1}^M \alpha_m = 1, \alpha_m \geq 0\} \times \prod_{m=1}^M \Omega_m.$$

For a given sample  $\mathbf{X} \equiv (X_1, \dots, X_n)$  from  $f(x; \theta_0)$ , the likelihood function is defined as

$$l(\theta; \mathbf{X}) \equiv \prod_{i=1}^n f(X_i; \theta) = \prod_{i=1}^n \left\{ \sum_{m=1}^M \alpha_m f_m(X_i; \mu_m, \sigma_m) \right\}.$$

Throughout this paper, we fix  $M$ , the number of components of mixture models. Let  $\mathcal{G}_m \subset \{f(x; \theta) \mid \theta \in \Theta\}$  denote the set of location-scale mixture densities which consist of no more than  $m$  components. For example, if a mixture density satisfies  $\alpha_{m+1} = \dots = \alpha_M = 0$ , then the density belongs to  $\mathcal{G}_m$ . Note that  $\mathcal{G}_M = \{f(x; \theta) \mid \theta \in \Theta\}$ .

Let  $\sigma_{(1)}$  and  $\sigma_{(M)}$  denote the minimum and the maximum values of the scale parameters:

$$\sigma_{(1)} \equiv \min_{1 \leq m \leq M} \sigma_m \quad , \quad \sigma_{(M)} \equiv \max_{1 \leq m \leq M} \sigma_m . \quad (2.1.1)$$

Let  $\{c_n\}_{n=1}^\infty$  and  $\{b_n\}_{n=1}^\infty$  denote sequences of positive reals which converge to zero. In our discussion, we use two constrained parameter space. Define  $\Theta_{c_n}, \Theta_{b_n}$  as follows:

$$\Theta_{c_n} \equiv \{\theta \in \Theta \mid \sigma_{(1)} \geq c_n\}, \quad \Theta_{b_n} \equiv \{\theta \in \Theta \mid \frac{\sigma_{(1)}}{\sigma_{(M)}} \geq b_n\}.$$

## 2.2 Regularity conditions

We introduce assumptions for the strong consistency of the maximum likelihood estimator. These assumptions are essentially the same as in Wald (1949), Redner (1981) and Tanaka and Takemura (2006).

Let  $\Gamma$  denote any compact subset of  $\Theta$ .

**Assumption 1.** For  $\theta \in \Theta$  and any positive real number  $\rho$ , let

$$f(x; \theta, \rho) \equiv \sup_{\text{dist}(\theta', \theta) \leq \rho} f(x; \theta'),$$

where  $\text{dist}(\theta', \theta)$  is a distance between  $\theta$  and  $\theta'$ . For each  $\theta \in \Gamma$  and sufficiently small  $\rho$ ,  $f(x; \theta, \rho)$  is measurable.

**Assumption 2.** For each  $\theta \in \Gamma$ , if  $\lim_{j \rightarrow \infty} \theta^{(j)} = \theta$ ,  $(\theta^{(j)} \in \Gamma)$  then  $\lim_{j \rightarrow \infty} f(x; \theta^{(j)}) = f(x; \theta)$  except on a set which is a null set and does not depend on the sequence  $\{\theta^{(j)}\}_{j=1}^\infty$ .

**Assumption 3.**

$$\int |\log f(x; \theta_0)| f(x; \theta_0) dx < \infty.$$

Furthermore, in Section 3, we impose Assumption 4 or 5 according to what type of penalty is made. If the penalty is imposed on the scale parameters themselves, then we impose Assumption 4. Alternatively, if the penalty is imposed on the ratios of the scale parameters, then we impose Assumption 5.

**Assumption 4.** There exist real constants  $v_0, v_1 > 0$  and  $\beta > 1$  such that

$$f_m(x; \mu_m = 0, \sigma_m = 1) \leq \min\{v_0, v_1 \cdot |x|^{-\beta}\}$$

for all  $m$ .

**Assumption 5.** *There exist real constants  $v_0, v_1 > 0$  and  $\beta > 2$  such that*

$$f_m(x; \mu_m = 0, \sigma_m = 1) \leq \min\{v_0, v_1 \cdot |x|^{-\beta}\}$$

*for all  $m$ .*

Note that Assumption 5 is stronger than Assumption 4. Therefore, if Assumption 1,2,3 and 5 hold, then Assumption 1,2,3 and 4 hold.

## 2.3 Strong Consistency

According to Redner (1981), we define strong consistency of estimators for mixture distributions by identifying the parameters whose densities are equal. Let

$$\Theta(\theta') \equiv \{\theta \in \Theta \mid f(x; \theta) = f(x; \theta') \text{ almost everywhere on } \mathbb{R}\}$$

Furthermore we abbreviate  $\Theta(\theta_0)$  as  $\Theta_0$ . Given  $U, V \subset \Theta$ , the distance between  $U$  and  $V$  are defined as

$$\text{dist}(U, V) \equiv \inf_{\theta \in U} \inf_{\theta' \in V} \text{dist}(\theta, \theta').$$

We now define strong consistency of estimators for mixture distributions as follows.

**Definition 1.** *An estimator  $\hat{\theta}_n$  is strongly consistent iff*

$$\text{Prob} \left( \lim_{n \rightarrow \infty} \text{dist}(\Theta(\hat{\theta}_n), \Theta_0) = 0 \right) = 1.$$

In this paper, two notations “ $\text{Prob}(A) = 1$ ” and “ $A, a.s.$ ” ( $A$  holds almost surely), will be used interchangeably.

## 3 Main results

### 3.1 Consistency of penalized maximum likelihood estimator when the penalty is imposed on the minimum of the ratios of the scale parameters

Now we define a penalized likelihood. Let  $\bar{r}_n(\cdot)$  denote a function on  $(0, 1]$  which satisfies the following assumption and is not identically equal to zero.

**Assumption 6.**  $\exists \bar{R} < \infty, \exists \bar{r} > 0, \exists \delta > 0, 0 \leq \exists \tilde{d} < 1$  such that

$$0 \leq \bar{r}_n(y) \leq \min \{ \bar{R}, \bar{r} \cdot y^{M+\delta} \cdot \exp(n^{\tilde{d}}) \}.$$

The Assumption 6 means that  $\bar{r}_n(y)$  is nonnegative, bounded in  $n$  and  $y$ , and converges to zero sufficiently fast as  $y$  tends to zero. Note that we can take a discontinuous function as  $\bar{r}_n(y)$ . In Corollary 1, we obtain the consistency of a constrained maximum likelihood estimator by using a discontinuous penalty function.

We define a penalty function  $1/r_n(\theta)$  or a reward function  $r_n(\theta)$  as

$$r_n(\theta) \equiv \bar{r}_n \left( \frac{\sigma_{(1)}}{\sigma_{(M)}} \right).$$

The penalized likelihood function is defined as  $g_n(\theta; \mathbf{X}) \equiv l(\theta; \mathbf{X}) \cdot r_n(\theta)$ . The penalized maximum likelihood estimator is defined as  $\hat{\theta}_{g_n} \equiv \operatorname{argsup}_{\theta \in \Theta} g_n(\theta; \mathbf{X})$ . As stated in Section 2, the likelihood  $l(\theta; \mathbf{X})$  may increase to infinity as  $\sigma_m$  decreases to zero. However, if the penalty  $1/r_n(\theta)$  increases to infinity or  $r_n(\theta)$  decreases to zero, the divergence of the likelihood may be avoided. This happens when a part of the scale parameters decreases to zero. If all the scale parameters decreases to zero, then the likelihood  $l(\theta; \mathbf{X})$  decreases to zero because a component with a very small scale parameter can only have a large contribution to a single observation at most. Therefore, the existence of the penalty term may prevent the positive divergence of the likelihood.

Let  $b_0 > 0$ . In this section, we take  $b_n$  as follows:

$$b_n = b_0 \cdot \exp(-n^{\tilde{d}})$$

We also assume the following conditions.

**Assumption 7.** *There exist a positive real  $r_{\Theta_0}$  and a positive integer  $N$  such that  $r_n(\theta) \geq r_{\Theta_0}$  for all  $\theta \in \Theta_0$  and  $n \geq N$ .*

If  $\bar{r}_n(y)$  is positive and unimodal or increasing, then  $r_n(\theta)$  satisfies Assumption 7. Assumption 7 guarantees that the penalized likelihood is nearly unaffected by the penalty term for  $\theta \in \Theta_0$  when sample size  $n$  is large.

**Assumption 8.** *There exist real constants  $d$ ,  $c_0$  and  $\Delta$  such that  $0 \leq \tilde{d} < d < 1$ ,  $c_0 > 0$ ,  $\Delta > 0$  and the following relation holds for all  $\theta \in \Theta_{c_n}$  and  $n \in \mathbb{N}$ :*

$$r_n(\theta) > (\sigma_{(1)})^M \quad \Rightarrow \quad \sigma_{(M)} < \frac{(\sigma_{(1)})^\Delta}{b_n},$$

where  $c_n = c_0 \cdot \exp(-n^d)$  and  $\Theta_{c_n} = \{\theta \in \Theta \mid \sigma_{(1)} \geq c_n\}$ .

Assumption 8 means that all the scale parameters of  $\theta \in \Theta_{c_n}$  are equally small if  $r_n(\theta) > (\sigma_{(1)})^M$ .

The assumptions for the penalties are not so restrictive. For example, if we set  $\bar{r}_n(y) = \bar{r} \cdot y^{\alpha-1} \cdot e^{n^{\tilde{d}}}$  and assume  $\alpha > M + 1$ , then  $\bar{r}_n(y)$  satisfies the Assumption 6 and  $r_n(\theta) = \bar{r}_n(\frac{\sigma_{(1)}}{\sigma_{(M)}})$  satisfies the Assumption 7 and 8.

Then the following theorem holds.

**Theorem 1.** *Suppose that  $\mathcal{G}_M$  satisfies the Assumption 1,2,3 and 5, and  $f(x; \theta_0) \in \mathcal{G}_M \setminus \mathcal{G}_{M-1}$ . Suppose that the penalty function  $r_n(\theta)$  satisfies the Assumption 6, 7 and 8. Then the penalized maximum likelihood estimator  $\hat{\theta}_{g_n}$  is strongly consistent.*

A proof of Theorem 1 is given in section 4.2.

As a corollary of Theorem 1, we can obtain the consistency of a constrained maximum likelihood estimator. Let us define the constrained maximum likelihood estimator restricted to  $\Theta_{b_n}$  as

$$\hat{\theta}_{b_n} \equiv \operatorname{argsup}_{\theta \in \Theta_{b_n}} l(\theta; \mathbf{X}).$$

If we put  $\bar{r}_n(y)$  and  $r_n(\theta)$  as

$$\bar{r}_n(y) = \begin{cases} 1 & (y \geq b_n) \\ 0 & (y < b_n) \end{cases}, \quad r_n(\theta) = \bar{r}_n\left(\frac{\sigma_{(1)}}{\sigma_{(M)}}\right) = \begin{cases} 1 & (\sigma_{(1)}/\sigma_{(M)} \geq b_n) \\ 0 & (\sigma_{(1)}/\sigma_{(M)} < b_n) \end{cases}, \quad (3.1.1)$$

then  $\hat{\theta}_{b_n}$  is equal to the penalized maximum likelihood estimator  $\hat{\theta}_{g_n} = \operatorname{argsup}_{\theta \in \Theta} g_n(\theta; \mathbf{X}) = \operatorname{argsup}_{\theta \in \Theta} l(\theta; \mathbf{X}) \cdot r_n(\theta)$ . If we take  $0 < \tilde{d} < 1$ , then  $r_n(\theta)$  given in (3.1.1) satisfies Assumption 6, 7 and 8. From this and Theorem 1, we obtain the following corollary.

**Corollary 1.** *Suppose that  $\mathcal{G}_M$  satisfies the Assumption 1,2,3 and 5, and  $f(x; \theta_0) \in \mathcal{G}_M \setminus \mathcal{G}_{M-1}$ . If we take  $0 < \tilde{d} < 1$ , then the constrained maximum likelihood estimator  $\hat{\theta}_{b_n}$  is strongly consistent.*

By Corollary 1, the problem stated in Hathaway (1985) is solved positively.

### 3.2 Consistency of penalized maximum likelihood estimator when the penalties are imposed on the scale parameters themselves

We also give a consistency result for the penalized maximum likelihood estimator in which the penalties are imposed on the scale parameters themselves. Let  $\bar{s}_n(\cdot)$  denote a function on  $(0, \infty)$  which satisfies the following assumptions.

**Assumption 9.**  $\bar{s}_n(y)$  is nonnegative, uniformly bounded and not identically equal to zero:

$$0 \leq \bar{s}_n(y) \leq \bar{S} < \infty \quad , \quad \sup_{y>0} \bar{s}_n(y) > 0.$$

**Assumption 10.**  $\bar{s}_n(y)$  converges to zero sufficiently fast as  $y$  tends to zero:

$$\exists \bar{s} > 0, 0 \leq \exists d < 1 \quad \text{s.t.} \quad 0 < \sup_{y>0} \frac{\bar{s}_n(y)}{y^M} \leq \bar{s} \cdot \exp(n^d)$$

Then we define a penalty function  $1/s_n(\theta)$  or reward function  $s_n(\theta)$  as follows:

$$s_n(\theta) \equiv \prod_{m=1}^M \bar{s}_n(\sigma_m).$$

The penalized likelihood function is defined as  $h_n(\theta; \mathbf{X}) \equiv l(\theta; \mathbf{X}) \cdot s_n(\theta)$ . The penalized maximum likelihood estimator is defined as  $\hat{\theta}_{h_n} \equiv \operatorname{argsup}_{\theta \in \Theta} h_n(\theta; \mathbf{X})$ .

We also assume the following condition.

**Assumption 11.** *There exist a positive real  $s_{\Theta_0}$  and a positive integer  $N$  such that  $s_n(\theta) \geq s_{\Theta_0}$  for all  $\theta \in \Theta_0$  and  $n \geq N$ .*

The assumptions for the penalty are not so restrictive. For example, if we set  $\bar{s}_n(y) = e^{-\frac{\beta}{y}} \cdot y^{-(\alpha+1)}$  and assume  $\alpha, \beta > 0$ , then  $\bar{s}_n(y)$  satisfies the Assumption 9 and 10, and  $s_n(\theta) = \prod_{m=1}^M \bar{s}_n(\sigma_m)$  satisfies the Assumption 11.

We now state the consistency of the penalized maximum likelihood estimator when the penalty is imposed on the scale parameters themselves.

**Theorem 2.** *Suppose that  $\mathcal{G}_M$  satisfies the Assumption 1,2,3 and 4, and  $f(x; \theta_0) \in \mathcal{G}_M \setminus \mathcal{G}_{M-1}$ . Suppose that the penalty function  $s_n(\theta)$  satisfies the Assumption 9,10 and 11. Then the penalized maximum likelihood estimator  $\hat{\theta}_{h_n}$  is strongly consistent.*

The statement of Theorem 2 is an extension of Corollary 1 of Ciuperca, Ridolfi, and Idier (2003). In their statement, penalties for the location parameters  $\mu_1, \dots, \mu_M$  may be required. This is because, in their proof, they use a compactification of the parameter space, but their penalized likelihood is not continuous over the compactified parameter space. For example, if  $\mu_1 \rightarrow \infty$ , then other components may still exist and hence their penalized likelihood may not tend to zero.

We give a proof of Theorem 2 in section 4.3.

## 4 Proofs

In this section, we prove Theorem 1 and 2. The organization of this section is as follows. In section 4.1, we state some lemmas needed for proving Theorem 1 and 2. Section 4.2 and 4.3 are devoted to the proof of Theorem 1 and 2 respectively.

### 4.1 Some lemmas

We state some lemmas needed for proving Theorem 1 and 2. Proofs of Lemma 4.1.1, 4.1.2, 4.1.5, 4.1.6, 4.1.7 and 4.1.8 are given in the longer version of Tanaka and Takemura (2006).

In Tanaka and Takemura (2006), we showed that when the constraint is appropriately imposed on the minimum of the scale parameters, the constrained maximum likelihood estimator is strongly consistent under regularity conditions. Let us define the



constrained maximum likelihood estimator restricted to  $\Theta_{c_n} = \{\theta \in \Theta \mid \sigma_{(1)} \geq c_n\}$  by  $\hat{\theta}_{c_n} \equiv \operatorname{argsup}_{\theta \in \Theta_{c_n}} l(\theta; \mathbf{X})$ .

**Lemma 4.1.1.** (Tanaka and Takemura (2006)) *Suppose that  $\mathcal{G}_M$  satisfies the Assumption 1,2,3 and 4, and  $f(x; \theta_0) \in \mathcal{G}_M \setminus \mathcal{G}_{M-1}$ . Let  $c_0 > 0$  and  $0 < d < 1$ . If  $c_n = c_0 \cdot \exp(-n^d)$ , then the constrained maximum likelihood estimator  $\hat{\theta}_{c_n}$  restricted to  $\Theta_{c_n}$  is strongly consistent.*

As in the case of uniform mixture in Tanaka and Takemura (2005), it is readily verified that if  $b_n$  decreases to zero faster than  $e^{-n}$ , then the consistency of the constrained maximum likelihood estimator fails. Therefore, the rate obtained in Lemma 4.1.1 is almost the lower bound of  $b_n$  which maintains the strong consistency.

Let

$$X_{n,1} \equiv \min \{X_1, \dots, X_n\} \quad , \quad X_{n,n} \equiv \max \{X_1, \dots, X_n\}.$$

**Lemma 4.1.2.** (Tanaka and Takemura (2006)) *Suppose that Assumption 4 is satisfied. For any real positive constants  $A_0 > 0, \zeta > 0$ , let*

$$A_n \equiv A_0 \cdot n^{\frac{2+\zeta}{\beta-1}}, \tag{4.1.1}$$

where  $\beta$  is defined by Assumption 4. Then

$$\operatorname{Prob}(X_{n,1} < -A_n \text{ or } X_{n,n} > A_n \text{ i.o.}) = 0.$$

where *i.o.* means “infinitely often”. By Lemma 4.1.2, we can bound the behavior of the minimum and the maximum of the sample with probability 1. In the following sections, we take  $A_0$  large enough to satisfy (4.2.16) and ignore the event  $\{X_{n,1} < -A_n \text{ or } X_{n,n} > A_n\}$ .

Let  $R_n(V)$  denote the number of observation which belong to a set  $V \subset \mathbb{R}$ :

$$R_n(V) \equiv \#\{X_i \mid X_i \in V, i = 1, \dots, n\}.$$

Let  $P_0(V)$  denote the probability of  $V \subset \mathbb{R}$  under the true density:

$$P_0(V) \equiv \int_V f(x; \theta_0) dx.$$

Let us consider an interval  $[\mu - w_n, \mu + w_n]$  with the center  $\mu$  and the length  $2w_n$ . If  $w_n = 0$ , then  $R_n([\mu - w_n, \mu + w_n])$  is clearly 0. In the following lemma, we state that if  $w_n$  decreases to zero faster than a power of  $1/n$ , then  $R_n([\mu - w_n, \mu + w_n]) < 2$  holds for every  $\mu \in \mathbb{R}$  with probability 1.

**Lemma 4.1.3.** *Suppose that Assumption 4 is satisfied. Let  $\{w_n\}_{n=1}^\infty$  be a sequence of real numbers which satisfies*

$$\lim_{n \rightarrow \infty} n^{3+\delta'} \cdot A_n \cdot w_n = 0, \tag{4.1.2}$$

where  $\delta' > 0$  and  $A_n$  is defined by (4.1.1). Then

$$\operatorname{Prob} \left( \sup_{\mu \in \mathbb{R}} R_n([\mu - w_n, \mu + w_n]) > 1 \text{ i.o.} \right) = 0.$$

**Proof:** From Lemma 4.1.2, we ignore the event  $\{X_{n,1} < -A_n \text{ or } X_{n,n} > A_n\}$ . Then

$$\sup_{\mu \in \mathbb{R}} R_n([\mu - w_n, \mu + w_n]) > 1 \quad \Leftrightarrow \quad \sup_{\mu \in [-A_n + w_n, A_n - w_n]} R_n([\mu - w_n, \mu + w_n]) > 1 \quad a.s. \quad (4.1.3)$$

Now we cover  $[-A_n, A_n]$  by short intervals of length  $4w_n$ . Let

$$\begin{aligned} I_1^{(n)} &\equiv [-A_n, -A_n + 4w_n], \quad I_2^{(n)} \equiv [-A_n + 2w_n, -A_n + 6w_n], \dots, \\ I_{k_n-1}^{(n)} &\equiv [-A_n + (k_n - 6) \cdot w_n, -A_n + (k_n - 2) \cdot w_n], \\ I_{k_n}^{(n)} &\equiv [-A_n + (k_n - 4) \cdot w_n, A_n], \end{aligned}$$

where  $k_n \equiv \min\{k \in \mathbb{N} \mid k \cdot (2w_n) > 2A_n\}$ . See Figure 1. Since any half-open interval of



Figure 1:  $I_1^{(n)}, I_2^{(n)}, \dots, I_{k_n}^{(n)}$

length  $2w_n$  in  $[-A_n, A_n]$  is covered by one of  $I_1^{(n)}, \dots, I_{k_n}^{(n)}$ , the following relation holds.

$$\sup_{\mu \in [-A_n + w_n, A_n - w_n]} R_n([\mu - w_n, \mu + w_n]) > 1 \quad \Rightarrow \quad 1 \leq \exists k \leq k_n, \quad R_n(I_k^{(n)}) > 1 \quad (4.1.4)$$

Let  $u_0 \equiv \sup_x f(x; \theta_0)$ . Because  $R_n(I_k^{(n)}) \sim \text{Bin}(n, P_0(I_k^{(n)}))$  and  $P_0(I_k^{(n)}) \leq 2w_n u_0$ , we obtain

$$\begin{aligned} \text{Prob} \left( 1 \leq \exists k \leq k_n, \quad R_n(I_k^{(n)}) > 1 \right) &\leq \sum_{k=1}^{k_n} \text{Prob} \left( R_n(I_k^{(n)}) > 1 \right) \\ &\leq k_n \cdot \left\{ \max_{1 \leq k \leq k_n} \text{Prob}(R_n(I_k^{(n)}) > 1) \right\} \\ &\leq \left( \frac{A_n}{w_n} + 1 \right) \cdot \sum_{k=2}^n \binom{n}{k} (2w_n u_0)^k (1 - 2w_n u_0)^{n-k} \\ &\leq \left( \frac{A_n}{w_n} + 1 \right) \cdot \sum_{k=2}^n \frac{n^k}{k!} (2w_n u_0)^k \\ &\leq \left( \frac{A_n}{w_n} + 1 \right) \cdot (2nw_n u_0)^2 \cdot \exp(2nw_n u_0). \end{aligned} \quad (4.1.5)$$

From (4.1.2), when we sum the right hand side of (4.1.5) over  $n$ , the resulting series converges. Hence by (4.1.3), (4.1.4), (4.1.5) and Borel-Cantelli lemma, we have

$$\text{Prob} \left( \sup_{\mu \in \mathbb{R}} R_n([\mu - w_n, \mu + w_n]) > 1 \quad i.o. \right) = 0.$$

□

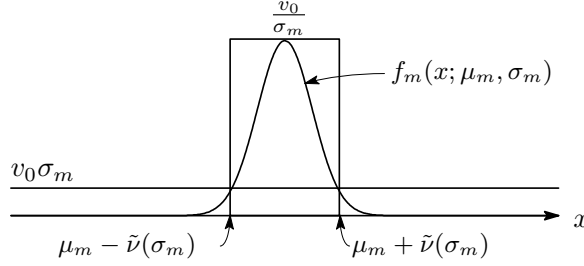


Figure 2: Each component is bounded by a step function.

Next we bound the component densities from above. For  $\beta > 2$ , define  $\tilde{\nu}(\sigma)$  as

$$\tilde{\nu}(\sigma) \equiv \left( \frac{v_1}{v_0} \right)^{\frac{1}{\beta}} \cdot \sigma^{1 - \frac{2}{\beta}}. \quad (4.1.6)$$

Let  $1_U(x)$  denote the indicator function of  $U \subset \mathbb{R}$ .

**Lemma 4.1.4.** *Suppose that Assumption 5 is satisfied. Then the following inequalities hold.*

$$f_m(x; \mu_m, \sigma_m) \leq \max \left\{ 1_{[\mu_m - \tilde{\nu}(\sigma_m), \mu_m + \tilde{\nu}(\sigma_m)]}(x) \cdot \frac{v_0}{\sigma_{(1)}}, v_0 \sigma_{(M)} \right\}, \quad 1 \leq m \leq M. \quad (4.1.7)$$

**Proof:** From Assumption 5, each component is bounded from above as

$$f_m(x; \mu_m, \sigma_m) \leq \max \left\{ 1_{[\mu_m - \tilde{\nu}(\sigma_m), \mu_m + \tilde{\nu}(\sigma_m)]}(x) \cdot \frac{v_0}{\sigma_m}, v_0 \sigma_m \right\}.$$

See figure 2. From this and (2.1.1), we obtain (4.1.7). □

Let  $E_0[\cdot]$  denote the expectation under the true parameter  $\theta_0$ .

**Lemma 4.1.5.** (Tanaka and Takemura (2006)) *Suppose that  $\mathcal{G}_M$  satisfies the Assumption 1, 2, 3 and 4, and  $f(x; \theta_0) \in \mathcal{G}_M \setminus \mathcal{G}_{M-1}$ . Then there exist real constants  $\kappa, \lambda > 0$  such that*

$$E_0 [\log \{f(x; \theta) + \kappa\}] + \lambda < E_0 [\log f(x; \theta_0)] \quad (4.1.8)$$

for all  $f(x; \theta) \in \mathcal{G}_{M-1}$ .

Fix arbitrary  $\kappa_0 > 0$ , which corresponds to  $\kappa$  in Lemma 4.1.5. For  $\beta > 1$ , define  $\nu(\sigma)$  as

$$\nu(\sigma) \equiv \left( \frac{v_1}{\kappa_0} \right)^{\frac{1}{\beta}} \cdot \sigma^{1-\frac{1}{\beta}}.$$

In a manner similar to the proof of Lemma 4.1.4, we can show the following lemma.

**Lemma 4.1.6.** (Tanaka and Takemura (2006)) *Suppose that Assumption 4 is satisfied. Then the following inequality holds.*

$$f_m(x; \mu_m, \sigma_m) \leq \max\{1_{[\mu_m-\nu(\sigma_m), \mu_m+\nu(\sigma_m)]}(x) \cdot \frac{v_0}{\sigma_m}, \kappa_0\}.$$

Lemma 4.1.6 bounds the tails of a density in a different way than in Lemma 4.1.4. On the one hand, in Lemma 4.1.4, the tails of a density is bounded by the value of scale parameter and Assumption 5 is needed because  $\beta$  should be larger than 2. On the other hand, in Lemma 4.1.6, the tails of a density is bounded by a constant and only Assumption 4 is needed. Lemma 4.1.4 will be used to prove Theorem 1. Lemma 4.1.6 will be used to prove Theorem 1 and Theorem 2. Therefore, Theorem 1 needs Assumption 5 which is stronger than Assumption 4.

Let  $\mathcal{K}$  be a subset of  $\{1, 2, \dots, M\}$  and let  $|\mathcal{K}|$  denote the number of elements in  $\mathcal{K}$ . Denote by  $\theta_{\mathcal{K}}$  a subvector of  $\theta \in \Theta$  consisting of the components in  $\mathcal{K}$ . Then the parameter space of subprobability measures consisting of the components in  $\mathcal{K}$  is

$$\bar{\Theta}_{\mathcal{K}} \equiv \{\theta_{\mathcal{K}} \mid \theta \in \Theta, \sum_{m \in \mathcal{K}} \alpha_m \leq 1\}.$$

Corresponding density and the set of subprobability densities are denoted by

$$f_{\mathcal{K}}(x; \theta_{\mathcal{K}}) \equiv \sum_{k \in \mathcal{K}} \alpha_k f_k(x; \mu_k, \sigma_k),$$

$$\mathcal{G}_{\mathcal{K}} \equiv \{f_{\mathcal{K}}(x; \theta_{\mathcal{K}}) \mid \theta_{\mathcal{K}} \in \bar{\Theta}_{\mathcal{K}}\}.$$

Then  $\mathcal{G}_K$ , the set of subprobability densities with no more than  $K$  components, can be represented as

$$\mathcal{G}_K \equiv \bigcup_{|\mathcal{K}| \leq K} \mathcal{G}_{\mathcal{K}} \quad (1 \leq K \leq M).$$

The following lemma follows from the bounded convergence theorem.

**Lemma 4.1.7.** (Tanaka and Takemura (2006)) *Let  $\Gamma_{\mathcal{K}}$  denote any compact subset of  $\bar{\Theta}_{\mathcal{K}}$ . For any real constant  $\kappa_0 \geq 0$  and any point  $\theta_{\mathcal{K}} \in \Gamma_{\mathcal{K}}$ , the following equality holds under Assumption 1 and 3.*

$$\lim_{\rho \rightarrow 0} E_0[\log\{f_{\mathcal{K}}(x; \theta_{\mathcal{K}}, \rho) + \kappa_0\}] = E_0[\log\{f_{\mathcal{K}}(x; \theta_{\mathcal{K}}) + \kappa_0\}].$$

The following lemma follows from lemma 4.1.7.

**Lemma 4.1.8.** (Tanaka and Takemura (2006)) *Let  $\kappa_0$  and  $\lambda_0$  be real constants which corresponds to  $\kappa$  and  $\lambda$  in Lemma 4.1.5. Let  $\Gamma_{\mathcal{K}}$  denote any compact subset of  $\bar{\Theta}_{\mathcal{K}}$ . Let  $\mathcal{B}(\theta_{\mathcal{K}}, \rho(\theta_{\mathcal{K}}))$  denote the open ball with center  $\theta_{\mathcal{K}}$  and radius  $\rho(\theta_{\mathcal{K}})$ . Suppose that Assumption 1 and 3 hold. Then  $\Gamma_{\mathcal{K}}$  can be covered by a finite number of balls  $\mathcal{B}(\theta_{\mathcal{K}}^{(1)}, \rho(\theta_{\mathcal{K}}^{(1)})), \dots, \mathcal{B}(\theta_{\mathcal{K}}^{(S)}, \rho(\theta_{\mathcal{K}}^{(S)}))$  such that*

$$E_0[\log \{f_{\mathcal{K}}(x; \theta_{\mathcal{K}}^{(s)}, \rho(\theta_{\mathcal{K}}^{(s)})) + \kappa_0\}] + \lambda_0 < E_0[\log f(x; \theta_0)] , \quad (s = 1, \dots, S) .$$

## 4.2 Proof of Theorem 1

First, we partition the parameter space  $\Theta$  into two sets. Then the proof of the strong consistency of the penalized maximum likelihood estimator is also partitioned into two parts. The proof for one set is obtained immediately by applying the result of Lemma 4.1.1.

### 4.2.1 Partitioning the parameter space

Let  $\tilde{d}$  be a constant defined by Assumption 6. Let  $d$  be a constant defined by Assumption 8. Define  $c_n = c_0 \cdot \exp(-n^d)$  and  $\Theta_{c_n} = \{\theta \in \Theta \mid \sigma_{(1)} \geq c_n\}$ . Then the parameter space  $\Theta$  is divided into two sets:

$$\Theta = \Theta_{c_n} \cup \Theta_{c_n}^C,$$

where  $\Theta_{c_n}^C = \{\theta \in \Theta \mid \sigma_{(1)} < c_n\}$  is the complement of  $\Theta_{c_n}$ . From Assumption 6, the reward term  $r_n(\theta)$  is bounded. Furthermore, Assumption 7 indicates that the asymptotic behavior is not affected by the penalty term around  $\Theta_0$ . Therefore the penalized maximum likelihood estimator over  $\Theta_{c_n}$  is strongly consistent by Lemma 4.1.1. If the maximum of the likelihood function over  $\Theta_{c_n}^C$  is very small, then the penalized maximum likelihood estimator over the whole parameter space  $\Theta$  is strongly consistent. This takes care of  $\Theta_{c_n}$  and from now on we consider the behavior of the penalized likelihood over  $\Theta_{c_n}^C$ .

Furthermore, we divide  $\Theta_{c_n}^C$  into two sets:

$$\Theta_{c_n}^C = \Phi_n \cup \Psi_n,$$

where

$$\Phi_n \equiv \{\theta \in \Theta_{c_n}^C \mid \frac{r_n(\theta)}{(\sigma_{(1)})^M} > 1\}, \quad (4.2.1)$$

$$\Psi_n \equiv \{\theta \in \Theta_{c_n}^C \mid \frac{r_n(\theta)}{(\sigma_{(1)})^M} \leq 1\}. \quad (4.2.2)$$

For  $\theta \in \Phi_n$ , all the scale parameters are very small. On the other hand,  $\theta \in \Psi_n$ , the penalty  $1/r_n(\theta)$  is very large and has large contribution relative to the likelihood. Therefore, intuitively, it seems that the maximum of the likelihood function over  $\Theta_{c_n}^C = \Phi_n \cup \Psi_n$  is very small. We are going to prove that this is true.

By the argument used in Wald (1949), in order to prove Theorem 1, it suffices to prove the following two equations.

$$\lim_{n \rightarrow \infty} \frac{\sup_{\theta \in \Phi_n} \{\prod_{i=1}^n f(X_i; \theta)\} \cdot r_n(\theta)}{\{\prod_{i=1}^n f(X_i; \theta_0)\} \cdot r_n(\theta_0)} = 0, \quad a.s. \quad (4.2.3)$$

$$\lim_{n \rightarrow \infty} \frac{\sup_{\theta \in \Psi_n} \{\prod_{i=1}^n f(X_i; \theta)\} \cdot r_n(\theta)}{\{\prod_{i=1}^n f(X_i; \theta_0)\} \cdot r_n(\theta_0)} = 0, \quad a.s. \quad (4.2.4)$$

#### 4.2.2 Proof of (4.2.3) for $\Phi_n$

By the law of large numbers, we have

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \log f(X_i; \theta_0) = E_0[\log f(x; \theta_0)], \quad a.s.$$

Furthermore, by Assumption 6 and 7, we obtain

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log r_n(\theta_0) = 0.$$

Therefore (4.2.3) is implied by

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \cdot \sup_{\theta \in \Phi_n} \left\{ \sum_{i=1}^n \log f(X_i; \theta) + \log r_n(\theta) \right\} < E_0[\log f(x; \theta_0)] \quad a.s. \quad (4.2.5)$$

Consequently, in order to prove (4.2.3), it suffices to prove (4.2.5).

From Assumption 8 and (4.2.2), we have

$$\sigma_{(1)} \leq \sigma_{(M)} < \frac{(\sigma_{(1)})^\Delta}{b_n}, \quad \theta \in \Phi_n, \quad (4.2.6)$$

where  $b_n = b_0 \cdot \exp(-n^{\tilde{d}})$  and the first inequality is derived from (2.1.1). Because  $\sigma_{(1)} < c_n = \exp(-n^d)$ , we obtain

$$\sigma_m < \exp(n^{\tilde{d}} - \Delta \cdot n^d), \quad 1 \leq m \leq M, \theta \in \Phi_n. \quad (4.2.7)$$

Note that  $0 \leq \tilde{d} < d < 1$ ,  $\Delta > 0$  by Assumption 8. Define

$$\tilde{J}(\theta) \equiv \bigcup_{m=1}^M [\mu_m - \tilde{\nu}(\sigma_m), \mu_m + \tilde{\nu}(\sigma_m)]. \quad (4.2.8)$$

Then the following lemma holds.

**Lemma 4.2.1.**

$$\text{Prob} \left( \sup_{\theta \in \Phi_n} R_n(\tilde{J}(\theta)) > M \quad i.o. \right) = 0. \quad (4.2.9)$$

**Proof:** We prove Lemma 4.2.1 by using Lemma 4.1.3. Let  $w_n = \tilde{\nu}(\exp(n^{\tilde{d}} - \Delta \cdot n^d))$ . Because (4.1.6) and  $\beta > 2$  by Assumption 5, the assumption (4.1.2) of Lemma 4.1.3 is satisfied. From (4.2.7) and (4.2.8), we have

$$\sup_{\theta \in \Phi_n} R_n(\tilde{J}(\theta)) > M \Rightarrow \sup_{\mu \in \mathbb{R}} R_n([\mu - w_n, \mu + w_n]) > 1$$

Therefore, by Lemma 4.1.3, we obtain (4.2.9).  $\square$

We now state the following inequality, in order to bound the likelihood.

**Lemma 4.2.2.** For  $\theta \in \Phi_n$ ,

$$\sum_{i=1}^n \log f(X_i; \theta) \leq R_n(\tilde{J}(\theta)) \cdot \log \frac{v_0}{\sigma_{(1)}} + R_n(\tilde{J}(\theta)^C) \cdot \log v_0 \sigma_{(M)}.$$

**Proof:** From Lemma 4.1.4 and (2.1.1), for  $\theta \in \Phi_n$ , we obtain

$$\begin{aligned} \sum_{i=1}^n \log f(X_i; \theta) &= \sum_{i=1}^n \log \left\{ \sum_{m=1}^M \alpha_m f_m(X_i; \mu_m, \sigma_m) \right\} \\ &\leq \sum_{i=1}^n \max_{m=1, \dots, M} \log f_m(X_i; \mu_m, \sigma_m) \\ &\leq \sum_{i=1}^n \max_{m=1, \dots, M} \left\{ \max\{1_{[\mu_m - \tilde{\nu}(\sigma_m), \mu_m + \tilde{\nu}(\sigma_m)]}(x) \cdot \log \frac{v_0}{\sigma_{(1)}}, \log v_0 \sigma_{(M)}\} \right\} \\ &= R_n(\tilde{J}(\theta)) \cdot \log \frac{v_0}{\sigma_{(1)}} + R_n(\tilde{J}(\theta)^C) \cdot \log v_0 \sigma_{(M)}. \end{aligned}$$

$\square$

By Lemma 4.2.2 and Assumption 6, we obtain for  $\theta \in \Phi_n$

$$\sum_{i=1}^n \log f(X_i; \theta) + \log r_n(\theta) \leq R_n(\tilde{J}(\theta)) \cdot \log \frac{v_0}{\sigma_{(1)}} + R_n(\tilde{J}(\theta)^C) \cdot \log v_0 \sigma_{(M)} + \log \bar{R}.$$

Furthermore, from (4.2.6), we have for  $\theta \in \Phi_n$

$$\begin{aligned} \sum_{i=1}^n \log f(X_i; \theta) + \log r_n(\theta) &\leq R_n(\tilde{J}(\theta)) \cdot \log \frac{v_0}{\sigma_{(1)}} + R_n(\tilde{J}(\theta)^C) \cdot \log \frac{v_0 (\sigma_{(1)})^\Delta}{b_n} + \log \bar{R} \\ &= \left( \Delta \cdot R_n(\tilde{J}(\theta)^C) - R_n(\tilde{J}(\theta)) \right) \cdot \log \sigma_{(1)} - R_n(\tilde{J}(\theta)^C) \cdot \log b_n + n \log v_0 + \log \bar{R}. \end{aligned}$$

Because  $b_n = b_0 \cdot e^{-n^{\tilde{d}}}$ , we obtain for  $\theta \in \Phi_n$

$$\begin{aligned} \sum_{i=1}^n \log f(X_i; \theta) + \log r_n(\theta) &\leq \left( \Delta \cdot R_n(\tilde{J}(\theta)^C) - R_n(\tilde{J}(\theta)) \right) \cdot \log \sigma_{(1)} + R_n(\tilde{J}(\theta)^C) \cdot (n^{\tilde{d}} - \log b_0) + n \log v_0 + \log \bar{R} \\ &\leq \left( \Delta \cdot R_n(\tilde{J}(\theta)^C) - R_n(\tilde{J}(\theta)) \right) \cdot \log \sigma_{(1)} + n \cdot (n^{\tilde{d}} + |-\log b_0| + \log v_0) + \log \bar{R}. \end{aligned} \tag{4.2.10}$$

By Lemma 4.2.1, we obtain

$$\begin{aligned}
1 &= \text{Prob} \left( \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \sup_{\theta \in \Phi_n} R_n(\tilde{J}(\theta)) \leq M \right) \\
&= \text{Prob} \left( \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \left\{ \left\{ \sup_{\theta \in \Phi_n} R_n(\tilde{J}(\theta)) \leq M \right\} \cap \left\{ \sup_{\theta \in \Phi_n} R_n(\tilde{J}(\theta)^C) \geq n - M \right\} \right\} \right) \\
&\leq \text{Prob} \left( \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \sup_{\theta \in \Phi_n} \left( \Delta \cdot R_n(\tilde{J}(\theta)^C) - R_n(\tilde{J}(\theta)) \right) \geq \Delta \cdot (n - M) - M \right) \leq 1.
\end{aligned} \tag{4.2.11}$$

From (4.2.11), the inequality  $\sup_{\theta \in \Phi_n} \Delta \cdot R_n(\tilde{J}(\theta)^C) - R_n(\tilde{J}(\theta)) \geq \Delta \cdot (n - M) - M$  holds almost surely except for finite number of  $n$ . Therefore, we ignore the event  $\sup_{\theta \in \Phi_n} \Delta \cdot R_n(\tilde{J}(\theta)^C) - R_n(\tilde{J}(\theta)) < \Delta \cdot (n - M) - M$ . Because  $\sigma_{(1)} \leq c_n = c_0 \cdot e^{-n^d}$  and (4.2.10), for all sufficiently large  $n$  such that  $c_n \leq 1$  and  $\Delta \cdot (n - M) - M \geq 0$  hold, we have

$$\begin{aligned}
&\sup_{\theta \in \Phi_n} \left\{ \sum_{i=1}^n \log f(X_i; \theta) + \log r_n(\theta) \right\} \\
&\leq (\Delta \cdot (n - M) - M) \cdot (-n^d + \log c_0) + n \cdot (n^{\bar{d}} + |-\log b_0| + \log v_0) + \log \bar{R} \quad a.s.
\end{aligned}$$

From Assumption 8, the first term of the right hand side of the above inequality is the main term and diverges to  $-\infty$  as  $n$  increases. Therefore, we obtain (4.2.5):

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \cdot \sup_{\theta \in \Phi_n} \left\{ \sum_{i=1}^n \log f(X_i; \theta) + \log r_n(\theta) \right\} = -\infty \quad a.s.$$

### 4.2.3 Proof of (4.2.4) for $\Psi_n$

The outline of the proof of (4.2.4) is as follows. First, we partition the parameter space  $\Psi_n$  into finite subsets  $\Psi_{n, \mathcal{K}, s}$  depending on the set of some parameter smaller than  $c_n$ . Then, by using Lemma 4.1.3, we can show that the components with  $\sigma_m < c_n$  do not contribute to the likelihood more than  $M$  data points and the contributions are canceled out by the penalty term. Therefore, from Lemma 4.1.4, 4.1.6 and 4.1.8, we obtain the following inequality for each  $\Psi_{n, \mathcal{K}, s}$ .

$$\limsup_{n \rightarrow \infty} \sup_{\theta \in \Psi_{n, \mathcal{K}, s}} \frac{1}{n} \sum_{i=1}^n \log f(X_i; \theta) + \log r_n(\theta) \leq E_0[\log f(x; \theta_0)] - \lambda_0, \quad a.s.$$

This leads to (4.2.4).



**Setting up constants** For  $\kappa, \lambda$  satisfying (4.1.8), let  $\kappa_0, \lambda_0$  be real constants such that

$$0 < 4\kappa_0 \leq \kappa \quad , \quad 0 < 4\lambda_0 \leq \lambda \quad , \quad \frac{v_0}{\kappa_0} > \max\{\sigma_{01}, \dots, \sigma_{0M}\}. \quad (4.2.12)$$

Note that  $4\kappa_0, 4\lambda_0$  also satisfy (4.1.8). Define

$$B \equiv \frac{v_0}{\kappa_0} \quad (4.2.13)$$

If  $\sigma_m \geq B$ , then the density of the  $m$ -th component is almost flat and makes little contribution to the likelihood. In the following argument, we partition the parameter space according to this property.

Because  $\{c_n\}$  is decreasing to zero, by replacing  $c_0$  by some  $c_n$  if necessary, we can assume without loss of generality that  $c_0$  is sufficiently small to satisfy the following conditions,

$$\begin{aligned} (v_0/c_0)^{\tilde{\beta}} &> e, \\ c_0 &< \min\{\sigma_{01}, \dots, \sigma_{0M}\}, \\ 3M \cdot u_0 \cdot 2\nu(c_0) \cdot |\log 2\kappa_0| &< \lambda_0, \\ 3 \cdot 2M \cdot u_0 \cdot \xi(v_0/c_0) \cdot \log(v_0/c_0) &< \lambda_0, \\ \kappa_0 &< \frac{v_0}{c_0(M+1)} \quad , \end{aligned} \quad (4.2.14)$$

where  $\tilde{\beta} \equiv (\beta - 1)/\beta$  and

$$\xi(y) \equiv 2 \cdot \left(\frac{v_1}{\kappa_0}\right)^{\frac{1}{\beta}} \cdot (v_0 \cdot (M+1))^{\tilde{\beta}} \cdot \left(\frac{1}{y}\right)^{\tilde{\beta}}. \quad (4.2.15)$$

Take  $A_0 > 0$  sufficiently large such that

$$P_0((-\infty, -A_0) \cup (A_0, \infty)) \cdot \log\left(\frac{v_0/c_0 + 3\kappa_0}{4\kappa_0}\right) < \lambda_0. \quad (4.2.16)$$

Let  $\mathcal{A}_0 \equiv (-\infty, -A_0) \cup (A_0, \infty)$  and  $A_n \equiv A_0 \cdot n^{\frac{2+\zeta}{\beta-1}}$  as in Lemma 4.1.2.

**Partitioning the parameter space** Partition  $\{1, \dots, M\}$  into disjoint subsets  $\mathcal{K}_{\sigma < c_n}$ ,  $\mathcal{K}_{c_n \leq \sigma < c_0}$ ,  $\mathcal{K}_{\sigma > B}$ ,  $\mathcal{K}_{|\mu| > A_0}$  and  $\mathcal{K}_R$ . For any given  $\mathcal{K}_{\sigma < c_n}$ ,  $\mathcal{K}_{c_n \leq \sigma < c_0}$ ,  $\mathcal{K}_{\sigma > B}$ ,  $\mathcal{K}_{|\mu| > A_0}$  and  $\mathcal{K}_R$ , we define a subset of  $\Psi_n$  by

$$\begin{aligned} \Psi_{n, \mathcal{K}} \equiv & \{ \theta \in \Psi_n \mid \sigma_m < c_n, (m \in \mathcal{K}_{\sigma < c_n}); \\ & c_n \leq \sigma_m < c_0, (m \in \mathcal{K}_{c_n \leq \sigma < c_0}); \\ & \sigma_m > B, (m \in \mathcal{K}_{\sigma > B}); \\ & c_0 \leq \sigma_m \leq B, |\mu_m| > A_0, (m \in \mathcal{K}_{|\mu| > A_0}); \\ & c_0 \leq \sigma_m \leq B, |\mu_m| \leq A_0, (m \in \mathcal{K}_R) \} \end{aligned}$$

The method of partitioning of the parameter space is the same as in Section 4.3.2 of Tanaka and Takemura (2006) except for  $\mathcal{K}_{\sigma < c_n}$ . We will show that the contributions of the components in  $\mathcal{K}_{\sigma < c_n}$  to the likelihood are canceled out by the penalty term.

As above, it suffices to prove that for each choice of disjoint subsets  $\mathcal{K}_{\sigma < c_n}$ ,  $\mathcal{K}_{c_n \leq \sigma < c_0}$ ,  $\mathcal{K}_{\sigma > B}$ ,  $\mathcal{K}_{|\mu| > A_0}$  and  $\mathcal{K}_R$

$$\lim_{n \rightarrow \infty} \frac{\sup_{\theta \in \Psi_{n, \mathcal{K}}} \prod_{i=1}^n f(X_i; \theta) \cdot r_n(\theta)}{\prod_{i=1}^n f(X_i; \theta_0) \cdot r_n(\theta_0)} = 0, \quad a.s.$$

We fix  $\mathcal{K}_{\sigma < c_n}$ ,  $\mathcal{K}_{c_n \leq \sigma < c_0}$ ,  $\mathcal{K}_{\sigma > B}$ ,  $\mathcal{K}_{|\mu| > A_0}$  and  $\mathcal{K}_R$  from now on.

Next we consider coverings of  $\bar{\Theta}_{\mathcal{K}_R}$ . The following lemma follows immediately from lemma 4.1.8 and compactness of  $\bar{\Theta}_{\mathcal{K}_R}$ .

**Lemma 4.2.3.** *Let  $\mathcal{B}(\theta, \rho(\theta))$  denote the open ball with center  $\theta$  and radius  $\rho(\theta)$ . Then  $\bar{\Theta}_{\mathcal{K}_R}$  can be covered by a finite number of balls  $\mathcal{B}(\theta_{\mathcal{K}_R}^{(1)}, \rho(\theta_{\mathcal{K}_R}^{(1)})), \dots, \mathcal{B}(\theta_{\mathcal{K}_R}^{(S)}, \rho(\theta_{\mathcal{K}_R}^{(S)}))$  such that*

$$E_0[\log \{f_{\mathcal{K}_R}(x; \theta_{\mathcal{K}_R}^{(s)}, \rho(\theta_{\mathcal{K}_R}^{(s)})) + \kappa_0\}] + \lambda_0 < E_0[\log f(x; \theta_0)], \quad (s = 1, \dots, S).$$

Based on lemma 4.2.3 we partition  $\Psi_{n, \mathcal{K}}$ . Recall that we denote by  $\theta_{\mathcal{K}}$  the subvector of  $\theta \in \Theta$  consisting of the components in  $\mathcal{K}$ . Define a subset of  $\Psi_{n, \mathcal{K}}$  by

$$\Psi_{n, \mathcal{K}, s} \equiv \{\theta \in \Psi_{n, \mathcal{K}} \mid \theta_{\mathcal{K}_R} \in \mathcal{B}(\theta_{\mathcal{K}_R}^{(s)}, \rho(\theta_{\mathcal{K}_R}^{(s)}))\}.$$

Then  $\Psi_{n, \mathcal{K}}$  is covered by  $\Psi_{n, \mathcal{K}, 1}, \dots, \Psi_{n, \mathcal{K}, S}$ :

$$\Psi_{n, \mathcal{K}} = \bigcup_{s=1}^S \Psi_{n, \mathcal{K}, s}.$$

Again it suffices to prove that for each choice of  $\mathcal{K}_{\sigma < c_n}$ ,  $\mathcal{K}_{c_n \leq \sigma < c_0}$ ,  $\mathcal{K}_{\sigma > B}$ ,  $\mathcal{K}_{|\mu| > A_0}$ ,  $\mathcal{K}_R$  and  $s$

$$\lim_{n \rightarrow \infty} \frac{\sup_{\theta \in \Psi_{n, \mathcal{K}, s}} \prod_{i=1}^n f(X_i; \theta) \cdot r_n(\theta)}{\prod_{i=1}^n f(X_i; \theta_0) \cdot r_n(\theta_0)} = 0, \quad a.s. \quad (4.2.17)$$

We fix  $\mathcal{K}_{\sigma < c_n}$ ,  $\mathcal{K}_{c_n \leq \sigma < c_0}$ ,  $\mathcal{K}_{\sigma > B}$ ,  $\mathcal{K}_{|\mu| > A_0}$ ,  $\mathcal{K}_R$  and  $s$  from now on. By Assumption 6, 7 and the law of large numbers, (4.2.17) is implied by

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \sup_{\theta \in \Psi_{n, \mathcal{K}, s}} \sum_{i=1}^n \log f(X_i; \theta) + \log r_n(\theta) < E_0[\log f(x; \theta_0)], \quad a.s. \quad (4.2.18)$$

Therefore it suffices to prove (4.2.18).

**Bounding the penalized likelihood by six terms** The outline of the rest of our proof is as follows. First, we bound the likelihood by four terms in Lemma 4.2.4. Next, we bound one term of the four terms obtained in Lemma 4.2.4 by three terms in Lemma 4.2.5. Finally, from Lemma 4.2.4 and Lemma 4.2.5, we bound the penalized likelihood by six terms in Lemma 4.2.6.

Define  $J_{\sigma < c_n}(\theta)$  as

$$J_{\sigma < c_n}(\theta) \equiv \bigcup_{m \in \mathcal{K}_{\sigma < c_n}} [\mu_m - \nu(\sigma_m), \mu_m + \nu(\sigma_m)]. \quad (4.2.19)$$

Let  $\mathcal{K}_{\sigma \geq c_n} = \{1, \dots, M\} \setminus \mathcal{K}_{\sigma < c_n}$ . Then the following lemma holds.

**Lemma 4.2.4.** For  $\theta \in \Psi_{n, \mathcal{K}, s}$ ,

$$\begin{aligned} & \sum_{i=1}^n \log f(X_i; \theta) \\ & \leq \sum_{i=1}^n \log \left\{ \sum_{m \in \mathcal{K}_{\sigma \geq c_n}} \alpha_m f_m(X_i; \theta_m) + \kappa_0 \right\} + R_n(J_{\sigma < c_n}(\theta)) \cdot \log \frac{v_0}{\kappa_0} \\ & \quad + R_n(J_{\sigma < c_n}(\theta)) \cdot \log \frac{1}{\sigma_{(1)}}. \end{aligned} \quad (4.2.20)$$

**Proof:** For  $\theta \in \Psi_{n, \mathcal{K}, s} \subset \Psi_n \subset \Theta_{c_n}^C$ , from Lemma 4.1.6, the following inequalities hold.

$$\begin{aligned} \sum_{i=1}^n \log f(X_i; \theta) &= \sum_{X_i \in J_{\sigma < c_n}(\theta)} \log f(X_i; \theta) + \sum_{X_i \in \mathbb{R} \setminus J_{\sigma < c_n}(\theta)} \log f(X_i; \theta) \\ &\leq R_n(J_{\sigma < c_n}(\theta)) \cdot \log \left\{ \max_{1 \leq m \leq M} \left( \frac{v_0}{\sigma_m} \right) \right\} \\ &\quad + \sum_{X_i \in \mathbb{R} \setminus J_{\sigma < c_n}(\theta)} \log \left\{ \sum_{m \in \mathcal{K}_{\sigma \geq c_n}} \alpha_m f_m(X_i; \theta_m) + \kappa_0 \right\} \\ &\leq \sum_{i=1}^n \log \left\{ \sum_{m \in \mathcal{K}_{\sigma \geq c_n}} \alpha_m f_m(X_i; \theta_m) + \kappa_0 \right\} - \sum_{X_i \in J_{\sigma < c_n}(\theta)} \log \kappa_0 \\ &\quad + R_n(J_{\sigma < c_n}(\theta)) \cdot \log \frac{v_0}{\sigma_{(1)}} \\ &\leq \sum_{i=1}^n \log \left\{ \sum_{m \in \mathcal{K}_{\sigma \geq c_n}} \alpha_m f_m(X_i; \theta_m) + \kappa_0 \right\} + R_n(J_{\sigma < c_n}(\theta)) \cdot \log \frac{v_0}{\kappa_0} \\ &\quad + R_n(J_{\sigma < c_n}(\theta)) \cdot \log \frac{1}{\sigma_{(1)}}. \end{aligned}$$

□

Define  $J_{c_n \leq \sigma < c_0}(\theta)$  as

$$J_{c_n \leq \sigma < c_0}(\theta) \equiv \bigcup_{m \in \mathcal{K}_{c_n \leq \sigma < c_0}} [\mu_m - \nu(\sigma_m), \mu_m + \nu(\sigma_m)].$$

For the first term of (4.2.20), the following lemma holds.

**Lemma 4.2.5.** *The following inequality holds for  $\theta \in \Psi_{n, \mathcal{K}, s}$ .*

$$\begin{aligned} \sum_{i=1}^n \log \left\{ \sum_{m \in \mathcal{K}_{\sigma \geq c_n}} \alpha_m f_m(X_i; \theta_m) + \kappa_0 \right\} &\leq \sum_{i=1}^n \log \{f_{\mathcal{K}_R}(X_i; \theta_{\mathcal{K}_R}, \rho(\theta_{\mathcal{K}_R})) + 4\kappa_0\} \\ &\quad + R_n(\mathcal{A}_0) \cdot \log \left( \frac{v_0/c_0 + 3\kappa_0}{4\kappa_0} \right) \\ &\quad + R_n(J_{c_n \leq \sigma < c_0}(\theta)) \cdot (-\log 2\kappa_0) \\ &\quad + \sum_{X_i \in J_{c_n \leq \sigma < c_0}(\theta)} \log \{f(X_i; \theta) + \kappa_0\}. \end{aligned} \quad (4.2.21)$$

**Proof:** Let  $\mathcal{K}_{\sigma \geq c_0} \equiv \{1, \dots, M\} \setminus \{\mathcal{K}_{\sigma < c_n} \cup \mathcal{K}_{c_n \leq \sigma < c_0}\}$  and  $\mathcal{K}_{c_n \leq \sigma \leq B} \equiv \{1, \dots, M\} \setminus \{\mathcal{K}_{\sigma < c_n} \cup \mathcal{K}_{c_n \leq \sigma < c_0} \cup \mathcal{K}_{\sigma > B}\}$ . For  $x \notin J_{c_n \leq \sigma < c_0}(\theta)$ ,  $f(x; \theta) \leq f_{\mathcal{K}_{\sigma \geq c_0}}(x; \theta_{\mathcal{K}_{\sigma \geq c_0}}) + \kappa_0$  holds. Therefore

$$\begin{aligned} \sum_{i=1}^n \log \{f(X_i; \theta) + \kappa_0\} &\leq \sum_{X_i \in J_{c_n \leq \sigma < c_0}(\theta)} \log \{f(X_i; \theta) + \kappa_0\} \\ &\quad + \sum_{X_i \notin J_{c_n \leq \sigma < c_0}(\theta)} \log \{f_{\mathcal{K}_{\sigma \geq c_0}}(x; \theta_{\mathcal{K}_{\sigma \geq c_0}}) + 2\kappa_0\} \\ &= \sum_{i=1}^n \log \{f_{\mathcal{K}_{\sigma \geq c_0}}(x; \theta_{\mathcal{K}_{\sigma \geq c_0}}) + 2\kappa_0\} \\ &\quad + \sum_{X_i \in J_{c_n \leq \sigma < c_0}(\theta)} \left[ \log \{f(X_i; \theta) + \kappa_0\} - \log \{f_{\mathcal{K}_{\sigma \geq c_0}}(x; \theta_{\mathcal{K}_{\sigma \geq c_0}}) + 2\kappa_0\} \right] \end{aligned} \quad (4.2.22)$$

Consider the second term on the right-hand side. We have

$$\begin{aligned} &\sum_{X_i \in J_{c_n \leq \sigma < c_0}(\theta)} \left[ \log \{f(X_i; \theta) + \kappa_0\} - \log \{f_{\mathcal{K}_{\sigma \geq c_0}}(x; \theta_{\mathcal{K}_{\sigma \geq c_0}}) + 2\kappa_0\} \right] \\ &\leq \sum_{X_i \in J_{c_n \leq \sigma < c_0}(\theta)} \log \{f(X_i; \theta) + \kappa_0\} - R_n(J_{c_n \leq \sigma < c_0}(\theta)) \cdot \log 2\kappa_0. \end{aligned}$$

This takes care of the third and the fourth term of (4.2.21). Now consider the first term on the right-hand side of (4.2.22). Because of (4.2.13), we obtain

$$\sum_{i=1}^n \log \{f_{\mathcal{K}_{\sigma \geq c_0}}(X_i; \theta_{\mathcal{K}_{\sigma \geq c_0}}) + 2\kappa_0\} \leq \sum_{i=1}^n \log \{f_{\mathcal{K}_{c_0 \leq \sigma \leq B}}(X_i; \theta_{\mathcal{K}_{c_0 \leq \sigma \leq B}}) + 3\kappa_0\}.$$

Note that  $\mathcal{A}_0 = \{x \in \mathbb{R} \mid |x| > A_0\}$  and  $\mathcal{K}_R = \mathcal{K}_{c_0 \leq \sigma \leq B} \setminus \mathcal{K}_{|\mu| > A_0}$ . For  $x \notin \mathcal{A}_0$ , we have

$$f_{\mathcal{K}_{|\mu| > A_0}}(x; \theta_{\mathcal{K}_{|\mu| > A_0}}) \leq \kappa_0.$$

Therefore we obtain

$$\begin{aligned} & \sum_{i=1}^n \log \left\{ f_{\mathcal{K}_{c_0 \leq \sigma \leq B}}(X_i; \theta_{\mathcal{K}_{c_0 \leq \sigma \leq B}}) + 3\kappa_0 \right\} \\ &= \sum_{X_i \notin \mathcal{A}_0} \log \left\{ f_{\mathcal{K}_{c_0 \leq \sigma \leq B}}(X_i; \theta_{\mathcal{K}_{c_0 \leq \sigma \leq B}}) + 3\kappa_0 \right\} + \sum_{X_i \in \mathcal{A}_0} \log \left\{ f_{\mathcal{K}_{c_0 \leq \sigma \leq B}}(X_i; \theta_{\mathcal{K}_{c_0 \leq \sigma \leq B}}) + 3\kappa_0 \right\} \\ &\leq \sum_{X_i \notin \mathcal{A}_0} \log \{f_{\mathcal{K}_R}(X_i; \theta_{\mathcal{K}_R}) + 4\kappa_0\} + \sum_{X_i \in \mathcal{A}_0} \log \left\{ f_{\mathcal{K}_{c_0 \leq \sigma \leq B}}(X_i; \theta_{\mathcal{K}_{c_0 \leq \sigma \leq B}}) + 3\kappa_0 \right\} \\ &= \sum_{i=1}^n \log \{f_{\mathcal{K}_R}(X_i; \theta_{\mathcal{K}_R}) + 4\kappa_0\} \\ &\quad + \sum_{X_i \in \mathcal{A}_0} \left[ \log \left\{ f_{\mathcal{K}_{c_0 \leq \sigma \leq B}}(X_i; \theta_{\mathcal{K}_{c_0 \leq \sigma \leq B}}) + 3\kappa_0 \right\} - \log \{f_{\mathcal{K}_R}(X_i; \theta_{\mathcal{K}_R}) + 4\kappa_0\} \right] \end{aligned} \tag{4.2.23}$$

Note that  $f_{\mathcal{K}_{c_0 \leq \sigma \leq B}}(x; \theta_{\mathcal{K}_{c_0 \leq \sigma \leq B}}) \leq v_0/c_0$  from lemma 4.1.4. Therefore

The r.h.s of (4.2.23)

$$\begin{aligned} &\leq \sum_{i=1}^n \log \{f_{\mathcal{K}_R}(X_i; \theta_{\mathcal{K}_R}) + 4\kappa_0\} + \sum_{X_i \in \mathcal{A}_0} [\log \{v_0/c_0 + 3\kappa_0\} - \log 4\kappa_0] \\ &\leq \sum_{i=1}^n \log \{f_{\mathcal{K}_R}(X_i; \theta_{\mathcal{K}_R}, \rho(\theta_{\mathcal{K}_R})) + 4\kappa_0\} + R_n(\mathcal{A}_0) \cdot \log \left( \frac{v_0/c_0 + 3\kappa_0}{4\kappa_0} \right). \end{aligned}$$

This takes care of the first and the second term of (4.2.21).  $\square$

By Lemma 4.2.4 and 4.2.5, the log likelihood function is bounded above as the following lemma.

**Lemma 4.2.6.** For  $\theta \in \Psi_{n, \mathcal{K}, s}$ ,

$$\begin{aligned}
& \sum_{i=1}^n \log f(X_i; \theta) + \log r_n(\theta) \\
& \leq \sum_{i=1}^n \log \{f_{\mathcal{K}_R}(X_i; \theta_{\mathcal{K}_R}, \rho(\theta_{\mathcal{K}_R})) + 4\kappa_0\} \\
& \quad + R_n(\mathcal{A}_0) \cdot \log \left( \frac{v_0/c_0 + 3\kappa_0}{4\kappa_0} \right) \\
& \quad + R_n(J_{c_n \leq \sigma < c_0}(\theta)) \cdot (-\log 2\kappa_0) \\
& \quad + \sum_{X_i \in J_{c_n \leq \sigma < c_0}(\theta)} \log \{f(X_i; \theta) + \kappa_0\} \\
& \quad + R_n(J_{\sigma < c_n}(\theta)) \cdot \log \frac{v_0}{\kappa_0} \\
& \quad + \left\{ R_n(J_{\sigma < c_n}(\theta)) \cdot \log \frac{1}{\sigma_{(1)}} + \log r_n(\theta) \right\}. \tag{4.2.24}
\end{aligned}$$

We bound the six terms of (4.2.24) in the following paragraphs.

**The first term** We begin by bounding the first term of (4.2.24). By lemma 4.1.8 and the strong law of large numbers, we have

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \log \{f_{\mathcal{K}_R}(X_i; \theta_{\mathcal{K}_R}, \rho(\theta_{\mathcal{K}_R})) + 4\kappa_0\} < E_0[\log f(x; \theta_0)] - 4\lambda_0, \quad a.s. \tag{4.2.25}$$

**The second term** By (4.2.16) and the strong law of large numbers, we have

$$\limsup_{n \rightarrow \infty} \frac{1}{n} R_n(\mathcal{A}_0) \cdot \log \left( \frac{v_0/c_0 + 3\kappa_0}{4\kappa_0} \right) < \lambda_0, \quad a.s. \tag{4.2.26}$$

**The third term and the fourth term** The third term and the fourth term of (4.2.24) can be bounded from above as follows:

$$\limsup_{n \rightarrow \infty} \sup_{\theta \in \Psi_{n, \mathcal{K}, s}} \frac{1}{n} R_n(J_{c_n \leq \sigma < c_0}(\theta)) \cdot |\log 2\kappa_0| \leq 3M \cdot u_0 \cdot 2\nu(c_0) \cdot |\log 2\kappa_0| < \lambda_0, \quad a.s. \tag{4.2.27}$$

$$\limsup_{n \rightarrow \infty} \sup_{\theta \in \Psi_{n, \mathcal{K}, s}} \frac{1}{n} \sum_{X_i \in J_{c_n \leq \sigma < c_0}(\theta)} \log \{f(X_i; \theta) + \kappa_0\} \leq \lambda_0, \quad a.s. \tag{4.2.28}$$

The proofs of the above inequalities are similar to the proofs of section 4.3.4 and 4.3.5 in the longer version of Tanaka and Takemura (2006), and are omitted.

**The fifth term and the sixth term** We now state the following lemma in order to bound the fifth term and sixth term of (4.2.24).

**Lemma 4.2.7.**

$$\text{Prob} \left( \sup_{\theta \in \Psi_{n,\mathcal{K},s}} R_n(J_{\sigma < c_n}(\theta)) > M \quad i.o. \right) = 0 \quad (4.2.29)$$

**Proof:** Let  $w_n = \nu(c_n) = \nu(\exp(n^{-d}))$ . Then (4.1.2), the assumption of Lemma 4.1.3, is satisfied. From (4.2.19), we have

$$\sup_{\theta \in \Psi_{n,\mathcal{K},s}} R_n(J_{\sigma < c_n}(\theta)) > M \quad \Rightarrow \quad \max_{\mu \in \mathbb{R}} R_n([\mu - w_n, \mu + w_n]) > 1$$

Therefore, by Lemma 4.1.3, we obtain (4.2.29).  $\square$

By Lemma 4.2.7 and the same argument in Section 4.2.2, we ignore the event  $R_n(J_{\sigma < c_n}(\theta)) > M$ . Then we have for  $\theta \in \Psi_{n,\mathcal{K},s}$  uniformly

$$\begin{aligned} R_n(J_{\sigma < c_n}(\theta)) \cdot \log \frac{v_0}{\kappa_0} + \left\{ R_n(J_{\sigma < c_n}(\theta)) \cdot \log \frac{1}{\sigma_{(1)}} + \log r_n(\theta) \right\} \\ \leq M \cdot \log \frac{v_0}{\kappa_0} + \log \frac{r_n(\theta)}{(\sigma_{(1)})^M}, \quad a.s. \end{aligned}$$

From (4.2.2), we obtain for  $\theta \in \Psi_{n,\mathcal{K},s}$

$$\begin{aligned} R_n(J_{\sigma < c_n}(\theta)) \cdot \log \frac{v_0}{\kappa_0} + \left\{ R_n(J_{\sigma < c_n}(\theta)) \cdot \log \frac{1}{\sigma_{(1)}} + \log r_n(\theta) \right\} \\ \leq M \cdot \log \frac{v_0}{\kappa_0}, \quad a.s. \end{aligned}$$

Because the right hand side of the above inequality is constant, we have

$$\begin{aligned} \limsup_{n \rightarrow \infty} \frac{1}{n} \cdot \sup_{\theta \in \Psi_{n,\mathcal{K},s}} \left[ R_n(J_{\sigma < c_n}(\theta)) \cdot \log \frac{v_0}{\kappa_0} + \left\{ R_n(J_{\sigma < c_n}(\theta)) \cdot \log \frac{1}{\sigma_{(1)}} + \log r_n(\theta) \right\} \right] \\ \leq 0 \quad a.s. \end{aligned} \quad (4.2.30)$$

**The end of the proof** Combining (4.2.24), (4.2.25), (4.2.26), (4.2.27), (4.2.28), (4.2.30), we obtain

$$\limsup_{n \rightarrow \infty} \sup_{\theta \in \Psi_{n,\mathcal{K},s}} \frac{1}{n} \sum_{i=1}^n \log f(X_i; \theta) + \log r_n(\theta) \leq E_0[\log f(x; \theta_0)] - \lambda_0, \quad a.s.$$

Therefore we obtain (4.2.18).

This completes the proof of Theorem 1.

### 4.3 Proof of Theorem 2

The outline of the proof of Theorem 2 is similar to the proof of Theorem 1.

**Partitioning the parameter space** Let  $d$  be a constant defined by Assumption 10. Define  $c_n = c_0 \cdot \exp(-n^d)$  and  $\Theta_{c_n} = \{\theta \in \Theta \mid \sigma_{(1)} \geq c_n\}$ . The parameter space  $\Theta$  is divided into two sets.

$$\Theta = \Theta_{c_n} \cup \Theta_{c_n}^C.$$

Because the asymptotic behavior is not affected by the penalty term, the penalized maximum likelihood estimator over  $\Theta_{c_n}$  is strongly consistent by Lemma 4.1.1. Therefore, it suffices to prove the following equation.

$$\lim_{n \rightarrow \infty} \frac{\sup_{\theta \in \Theta_{c_n}^C} \{\prod_{i=1}^n f(X_i; \theta)\} \cdot s_n(\theta)}{\{\prod_{i=1}^n f(X_i; \theta_0)\} \cdot s_n(\theta_0)} = 0, \quad a.s. \quad (4.3.1)$$

**Setting up constants** We set up some constants as in section 4.2.3.

Let  $\kappa_0, \lambda_0$  be real constants such that (4.2.12) holds. We can assume without loss of generality that  $c_0$  is sufficiently small to satisfy the equations (4.2.14). Take  $A_0 > 0$  sufficiently large such that (4.2.16) holds. Let  $\mathcal{A}_0 \equiv (-\infty, -A_0) \cup (A_0, \infty)$  and  $A_n \equiv A_0 \cdot n^{\frac{2+\zeta}{\beta-1}}$  as in lemma 4.1.2. Remember that  $\tilde{\beta} = (\beta - 1)/\beta$ , and  $B$  and  $\xi$  are defined in (4.2.13) and (4.2.15) respectively.

**Partitioning the parameter space** Partition  $\{1, \dots, M\}$  into disjoint subsets  $\mathcal{K}_{\sigma < c_n}$ ,  $\mathcal{K}_{c_n \leq \sigma < c_0}$ ,  $\mathcal{K}_{\sigma > B}$ ,  $\mathcal{K}_{|\mu| > A_0}$  and  $\mathcal{K}_R$ . For any given  $\mathcal{K}_{\sigma < c_n}$ ,  $\mathcal{K}_{c_n \leq \sigma < c_0}$ ,  $\mathcal{K}_{\sigma > B}$ ,  $\mathcal{K}_{|\mu| > A_0}$  and  $\mathcal{K}_R$ , we define a subset of  $\Theta_{c_n}^C$  by

$$\begin{aligned} \Theta_{c_n, \mathcal{K}}^C \equiv & \{ \theta \in \Theta_{c_n}^C \mid \sigma_m < c_n, (m \in \mathcal{K}_{\sigma < c_n}); \\ & c_n \leq \sigma_m < c_0, (m \in \mathcal{K}_{c_n \leq \sigma < c_0}); \\ & \sigma_m > B, (m \in \mathcal{K}_{\sigma > B}); \\ & c_0 \leq \sigma_m \leq B, |\mu_m| > A_0, (m \in \mathcal{K}_{|\mu| > A_0}); \\ & c_0 \leq \sigma_m \leq B, |\mu_m| \leq A_0, (m \in \mathcal{K}_R) \} \end{aligned}$$

As above, it suffices to prove that for each choice of disjoint subsets  $\mathcal{K}_{\sigma < c_n}$ ,  $\mathcal{K}_{c_n \leq \sigma < c_0}$ ,  $\mathcal{K}_{\sigma > B}$ ,  $\mathcal{K}_{|\mu| > A_0}$  and  $\mathcal{K}_R$

$$\lim_{n \rightarrow \infty} \frac{\sup_{\theta \in \Theta_{c_n, \mathcal{K}}^C} \{\prod_{i=1}^n f(X_i; \theta)\} \cdot s_n(\theta)}{\{\prod_{i=1}^n f(X_i; \theta_0)\} \cdot s_n(\theta_0)} = 0, \quad a.s.$$

We fix  $\mathcal{K}_{\sigma < c_n}$ ,  $\mathcal{K}_{c_n \leq \sigma < c_0}$ ,  $\mathcal{K}_{\sigma > B}$ ,  $\mathcal{K}_{|\mu| > A_0}$  and  $\mathcal{K}_R$  from now on.

Next we consider coverings of  $\Theta_{\mathcal{K}_R}$ . The following lemma follows immediately from lemma 4.1.8 and compactness of  $\bar{\Theta}_{\mathcal{K}_R}$ .



**Lemma 4.3.1.** *Let  $\mathcal{B}(\theta, \rho(\theta))$  denote the open ball with center  $\theta$  and radius  $\rho(\theta)$ . Then  $\bar{\Theta}_{\mathcal{K}_R}$  can be covered by a finite number of balls  $\mathcal{B}(\theta_{\mathcal{K}_R}^{(1)}, \rho(\theta_{\mathcal{K}_R}^{(1)})), \dots, \mathcal{B}(\theta_{\mathcal{K}_R}^{(S)}, \rho(\theta_{\mathcal{K}_R}^{(S)}))$  such that*

$$E_0[\log \{f_{\mathcal{K}_R}(x; \theta_{\mathcal{K}_R}^{(s)}, \rho(\theta_{\mathcal{K}_R}^{(s)})) + \kappa_0\}] + \lambda_0 < E_0[\log f(x; \theta_0)] , \quad (s = 1, \dots, S) .$$

Based on lemma 4.3.1, we partition  $\Theta_{c_n, \mathcal{K}}^C$ . Define a subset of  $\Theta_{c_n, \mathcal{K}}^C$  by

$$\Theta_{c_n, \mathcal{K}, s}^C \equiv \{\theta \in \Theta_{c_n, \mathcal{K}}^C \mid \theta_{\mathcal{K}_R} \in \mathcal{B}(\theta_{\mathcal{K}_R}^{(s)}, \rho(\theta_{\mathcal{K}_R}^{(s)}))\}.$$

Then  $\Theta_{c_n, \mathcal{K}}^C$  is covered by  $\Theta_{c_n, \mathcal{K}, 1}^C, \dots, \Theta_{c_n, \mathcal{K}, S}^C$  :

$$\Theta_{c_n, \mathcal{K}}^C = \bigcup_{s=1}^S \Theta_{c_n, \mathcal{K}, s}^C .$$

Again it suffices to prove that for each choice of  $\mathcal{K}_{\sigma < c_n}$ ,  $\mathcal{K}_{c_n \leq \sigma < c_0}$ ,  $\mathcal{K}_{\sigma > B}$ ,  $\mathcal{K}_{|\mu| > A_0}$ ,  $\mathcal{K}_R$  and  $s$

$$\lim_{n \rightarrow \infty} \frac{\sup_{\theta \in \Theta_{c_n, \mathcal{K}, s}^C} \{\prod_{i=1}^n f(X_i; \theta)\} \cdot s_n(\theta)}{\{\prod_{i=1}^n f(X_i; \theta_0)\} \cdot s_n(\theta_0)} = 0, \quad a.s.$$

We fix  $\mathcal{K}_{\sigma < c_n}$ ,  $\mathcal{K}_{c_n \leq \sigma < c_0}$ ,  $\mathcal{K}_{\sigma > B}$ ,  $\mathcal{K}_{|\mu| > A_0}$ ,  $\mathcal{K}_R$  and  $s$  from now on.

By Assumption 9, 11 and law of large numbers, (4.3.1) is implied by

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \cdot \sup_{\theta \in \Theta_{c_n, \mathcal{K}, s}^C} \left\{ \sum_{i=1}^n \log f(X_i; \theta) + \log s_n(\theta) \right\} < E_0[\log f(x; \theta_0)] \quad a.s. \quad (4.3.2)$$

We prove (4.3.2) in the following paragraphs.

**Bounding the penalized likelihood function by six terms** Define  $J_{\sigma < c_n}(\theta)$  and  $J_{c_n \leq \sigma < c_0}$  as

$$\begin{aligned} J_{\sigma < c_n}(\theta) &\equiv \bigcup_{m \in \mathcal{K}_{\sigma < c_n}} [\mu_m - \nu(\sigma_m), \mu_m + \nu(\sigma_m)]. \\ J_{c_n \leq \sigma < c_0}(\theta) &\equiv \bigcup_{m \in \mathcal{K}_{c_n \leq \sigma < c_0}} [\mu_m - \nu(\sigma_m), \mu_m + \nu(\sigma_m)]. \end{aligned} \quad (4.3.3)$$

The following lemma can be proved by a method similar to the proof of Lemma 4.2.6.

**Lemma 4.3.2.** For  $\theta \in \Theta_{c_n, \mathcal{K}, s}^C$ ,

$$\begin{aligned}
& \sum_{i=1}^n \log f(X_i; \theta) + \log s_n(\theta) \\
& \leq \sum_{i=1}^n \log \{f_{\mathcal{K}_R}(X_i; \theta_{\mathcal{K}_R}, \rho(\theta_{\mathcal{K}_R})) + 4\kappa_0\} \\
& \quad + R_n(\mathcal{A}_0) \cdot \log \left( \frac{v_0/c_0 + 3\kappa_0}{4\kappa_0} \right) \\
& \quad + R_n(J_{c_n \leq \sigma < c_0}(\theta)) \cdot (-\log 2\kappa_0) \\
& \quad + \sum_{X_i \in J_{c_n \leq \sigma < c_0}(\theta)} \log \{f(X_i; \theta) + \kappa_0\} \\
& \quad + R_n(J_{\sigma < c_n}(\theta)) \cdot \log \frac{v_0}{\kappa_0} \\
& \quad + \left\{ R_n(J_{\sigma < c_n}(\theta)) \cdot \log \frac{1}{\sigma_{(1)}} + \log s_n(\theta) \right\}. \tag{4.3.4}
\end{aligned}$$

We bound the six terms of (4.3.4) in the following paragraphs.

**The first term** We begin by bounding the first term of (4.3.4). By lemma 4.1.8 and the strong law of large numbers, we have

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \log \{f_{\mathcal{K}_R}(X_i; \theta_{\mathcal{K}_R}, \rho(\theta_{\mathcal{K}_R})) + 4\kappa_0\} < E_0[\log f(x; \theta_0)] - 4\lambda_0, \quad a.s. \tag{4.3.5}$$

**The second term** By (4.2.16) and the strong law of large numbers, we have

$$\lim_{n \rightarrow \infty} \frac{1}{n} R_n(\mathcal{A}_0) \cdot \log \left( \frac{v_0/c_0 + 3\kappa_0}{4\kappa_0} \right) < \lambda_0, \quad a.s. \tag{4.3.6}$$

**The third term and the fourth term** The third term and fourth term of (4.3.4) can be bounded from above as follows:

$$\limsup_{n \rightarrow \infty} \sup_{\theta \in \Theta_{c_n, \mathcal{K}, s}^C} \frac{1}{n} R_n(J_{c_n \leq \sigma < c_0}(\theta)) \cdot |\log 2\kappa_0| \leq 3M \cdot u_0 \cdot 2\nu(c_0) \cdot |\log 2\kappa_0| < \lambda_0, \quad a.s. \tag{4.3.7}$$

$$\limsup_{n \rightarrow \infty} \sup_{\theta \in \Theta_{c_n, \mathcal{K}, s}^C} \frac{1}{n} \sum_{X_i \in J_{c_n \leq \sigma < c_0}(\theta)} \log \{f(X_i; \theta) + \kappa_0\} \leq \lambda_0, \quad a.s. \tag{4.3.8}$$

The proofs of the above inequalities are similar to the proofs of section 4.3.4 and 4.3.5 in longer version of Tanaka and Takemura (2006), and are omitted.

**The fifth term** We now state the Lemma 4.3.3 in order to bound the fifth term of (4.3.4).

**Lemma 4.3.3.**

$$\text{Prob} \left( \sup_{\theta \in \Theta_{c_n, \mathcal{K}, s}^C} R_n(J_{\sigma < c_n}(\theta)) > M \quad i.o. \right) = 0 \quad (4.3.9)$$

**Proof:** Let  $w_n = \nu(c_n) = \nu(\exp(n^{-d}))$ . Then (4.1.2), the assumption of Lemma 4.1.3, is satisfied. From (4.3.3), we have

$$\sup_{\theta \in \Theta_{c_n, \mathcal{K}, s}^C} R_n(J_{\sigma < c_n}(\theta)) > M \quad \Rightarrow \quad \max_{\mu \in \mathbb{R}} R_n([\mu - w_n, \mu + w_n]) > 1$$

Therefore, by Lemma 4.1.3, we obtain (4.3.9).  $\square$

By Lemma 4.3.3 and the same argument in Section 4.2.2, we ignore the event  $R_n(J_{\sigma < c_n}(\theta)) > M$ . Then we have

$$\sup_{\theta \in \Theta_{c_n, \mathcal{K}, s}^C} R_n(J_{\sigma < c_n}(\theta)) \cdot \log \frac{v_0}{\kappa_0} \leq M \cdot \log \frac{v_0}{\kappa_0} \quad a.s.$$

Therefore, we obtain for  $\theta \in \Theta_{c_n, \mathcal{K}, s}^C$

$$\lim_{n \rightarrow \infty} \sup_{\theta \in \Theta_{c_n, \mathcal{K}, s}^C} \frac{1}{n} \cdot R_n(J_{\sigma < c_n}(\theta)) \cdot \log \frac{v_0}{\kappa_0} = 0 \quad a.s. \quad (4.3.10)$$

**The sixth term** From Lemma 4.3.3 and the same argument in Section 4.2.2, we have for  $\theta \in \Theta_{c_n, \mathcal{K}, s}^C$  uniformly

$$R_n(J_{\sigma < c_n}(\theta)) \cdot \log \frac{1}{\sigma_{(1)}} + \log s_n(\theta) \leq \log \frac{s_n(\theta)}{(\sigma_{(1)})^M} \quad a.s.$$

Furthermore, from Assumption 9 and 10, we have

$$\frac{s_n(\theta)}{(\sigma_{(1)})^M} = \frac{\bar{s}_n(\sigma_{(1)})}{(\sigma_{(1)})^M} \cdot \prod_{m=2}^M \bar{s}_n(\sigma_{(m)}) \leq \bar{S}^{M-1} \cdot \bar{s} \cdot \exp(n^d).$$

Note that  $0 \leq d < 1$ . Therefore we obtain for  $\theta \in \Theta_{c_n, \mathcal{K}, s}^C$

$$\lim_{n \rightarrow \infty} \frac{1}{n} \cdot \left\{ R_n(J_{\sigma < c_n}(\theta)) \cdot \log \frac{1}{\sigma_{(1)}} + \log s_n(\theta) \right\} = 0 \quad a.s. \quad (4.3.11)$$

**The end of the proof** From (4.3.5), (4.3.6), (4.3.7), (4.3.8), (4.3.10), (4.3.11), and Lemma 4.3.2, we have

$$\lim_{n \rightarrow \infty} \sup_{\theta \in \Theta_{c_n, \mathcal{K}, s}^C} \frac{1}{n} \left\{ \sum_{i=1}^n \log f(X_i; \theta) + \log s_n(\theta) \right\} < E_0[\log f(x; \theta_0)] - \lambda_0 \quad a.s.$$

Therefore we obtain (4.3.2).

This completes the proof of Theorem 2.

## 5 Conclusion

In location-scale mixture distributions, we have shown the consistency results for the two types of penalized maximum likelihood estimators. In Corollary 1, an open problem mentioned in Hathaway (1985), McLachlan and Peel (2000) has been solved positively as follows:

- It is possible to let the lower bound  $b$  of the ratios of variances decrease to zero as the sample size  $n$  increases to infinity while maintaining consistency.
- If the rate of convergence of  $b$  is slower than  $\exp(-n^{\tilde{d}})$  where  $\tilde{d}$  is a constant such that  $0 < \tilde{d} < 1$ , then the maximum likelihood estimator is strongly consistent under the constraint  $\min_{m,m'} \frac{\sigma_m}{\sigma_{m'}} \geq b$ .

The assumptions for the penalties given in section 3.1 or section 3.2 are not so restrictive. Note that the penalty does not have to depend on the sample size  $n$ . For example, if we set  $\bar{r}_n(y) = \bar{r} \cdot y^{\alpha-1}$  and assume  $\alpha > M + 1$ , then  $\bar{r}_n(y)$  satisfies the Assumption 6 and  $r_n(\theta) = \bar{r}_n(\frac{\theta_{(1)}}{\theta_{(M)}})$  satisfies the Assumption 7 and 8. The penalized likelihood  $g_n(\theta; \mathbf{X})$  corresponds to the posterior likelihood when we adopt a beta distribution as the prior of the minimum of the ratios of the scale parameters. Furthermore, if we set  $\bar{s}_n(y) = e^{-\frac{\beta}{y}} \cdot y^{-(\alpha+1)}$  and assume  $\alpha, \beta > 0$ , then  $\bar{s}_n(y)$  satisfies the Assumption 9 and 10, and  $s_n(\theta) = \prod_{m=1}^M \bar{s}_n(\sigma_m)$  satisfies the Assumption 11. The penalized likelihood  $h_n(\theta; \mathbf{X})$  corresponds to the posterior likelihood when we adopt inverse gamma distributions as the priors of the scale parameters.

From Theorem 1 and 2, we can easily show that the consistency of penalized likelihood estimator holds even when restrictions on either the location or scale parameters exist. If we know that the true parameter is in the restricted parameter space and the assumptions hold, then the consistency of the penalized maximum likelihood estimator holds by setting  $r_n(\theta) = 0$  or  $s_n(\theta) = 0$  for all  $n$  outside the restricted parameter space. For example, suppose one considers a uniform mixture distributions under the assumption that the data is non-negative. Theorem 1 and 2 are applicable to this model.

## References

- CIUPERCA, G., A. RIDOLFI, AND J. IDIER (2003): “Penalized maximum likelihood estimator for normal mixtures,” *Scand. J. Statist.*, 30(1), 45–49.
- DAY, N. (1969): “Estimating the components of a mixture of normal distributions,” *Biometrika*, 56(2), 463–474.
- GEMAN, S., AND C.-R. HWANG (1982): “Nonparametric maximum likelihood estimation by the method of sieves,” *Ann. Statist.*, 10(2), 401–414.
- GENOVESE, C. R., AND L. WASSERMAN (2000): “Rates of convergence for the Gaussian mixture sieve,” *Ann. Statist.*, 28(4), 1105–1127.

- GHOSAL, S., AND A. W. VAN DER VAART (2001): “Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities,” *Ann. Statist.*, 29(5), 1233–1263.
- GRENANDER, U. (1981): *Abstract inference*. Wiley, New York.
- HATHAWAY, R. J. (1985): “A constrained formulation of maximum-likelihood estimation for normal mixture distributions,” *Ann. Statist.*, 13(2), 795–800.
- MCLACHLAN, G. J., AND D. PEEL (2000): *Finite mixture models*. Wiley, New York.
- REDNER, R. (1981): “Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions,” *Ann. Statist.*, 9(1), 225–228.
- TANAKA, K., AND A. TAKEMURA (2005): “Strong consistency of MLE for finite uniform mixtures when the scale parameters are exponentially small,” *Ann. Inst. Statist. Math.*, 57(1), 1–19.
- (2006): “Strong consistency of MLE for finite mixtures of location-scale distributions when the scale parameters are exponentially small,” *Bernoulli*, 12(6), 1003–1017, (Longer version is available at <http://arXiv.org/abs/math.ST/0605148>).
- WALD, A. (1949): “Note on the consistency of the maximum likelihood estimate,” *Ann. Math. Statist.*, 20, 595–601.